

# caBIG Data Analysis & Statistical Tools Meeting

March 4, 2005

## Roll Call

## Discussion

- ◆ Vocabulary Survey Form – Dr. Leo Cheung
- ◆ CGH – Ping Liang, Chris Kingsley
- ◆ Normalization / local storage issues

## Discussions

- ◆ Project maturity stratification
- ◆ Workflow discussion (Patrick McConnell)
  - Use cases received?
- ◆ caArray discussion
- ◆ Available online:  
[http://cabig.nci.nih.gov/workspaces/ICR/Meetings/SIGs/Data\\_analysis\\_stats](http://cabig.nci.nih.gov/workspaces/ICR/Meetings/SIGs/Data_analysis_stats)

Dr. Leo Cheung

## Comparative Genomic Hybridization (CGH)

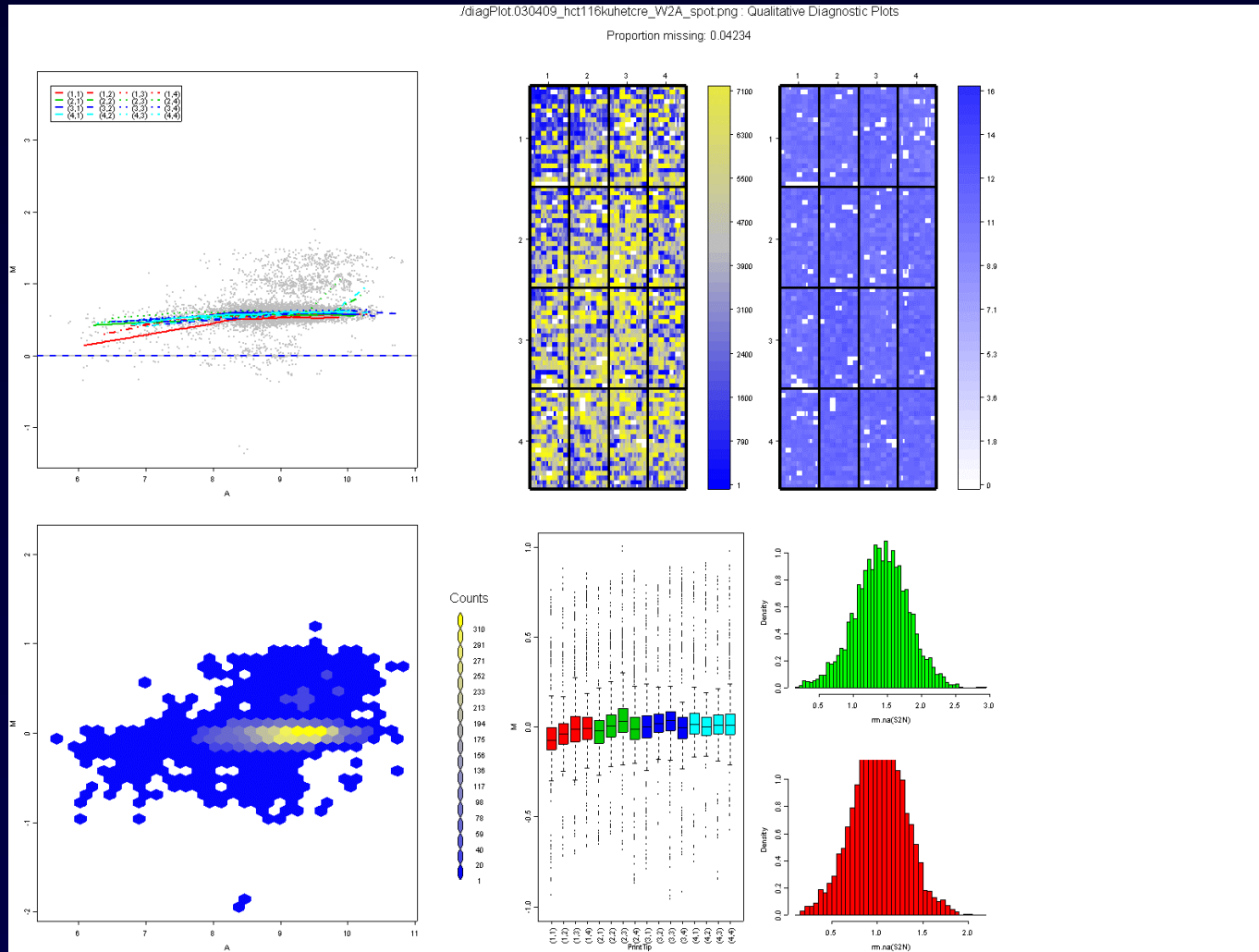
- ◆ A measure of genomic copy number changes across the genome.
  - ◆ Most solid tumors display some copy number changes
  - ◆ Copy number changes have been shown to affect transcription levels for large numbers of genes (Pollack et al, PNAS 99 (2002)12963-12968)
  - ◆ Specific copy number abnormalities are associated with certain tumor types, and some have been shown to correlate with clinical variables

Dr. Ping Liang – Roswell Park

## aCGH package (Bioconductor)

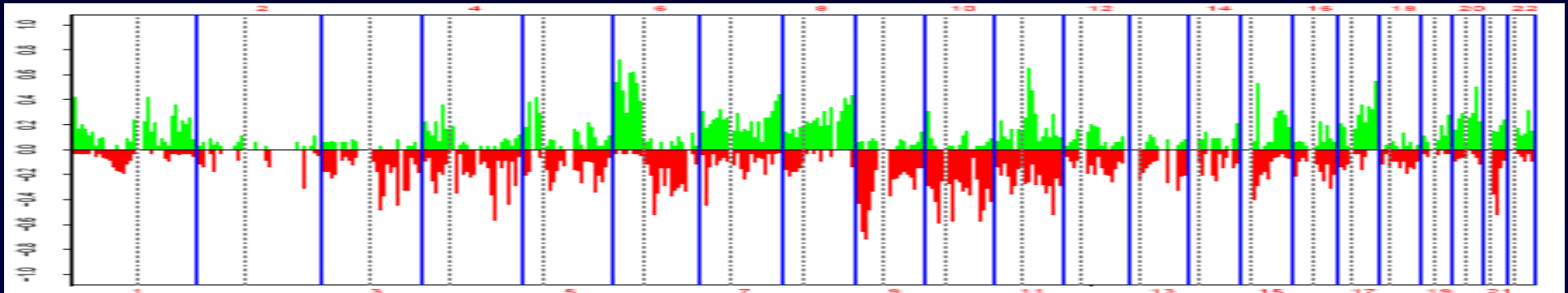
- ◆ Developed by Jane Fridlyand (UCSF) and Peter Dimitrov(UCB)
- ◆ Provides environment for visualization and analysis of DNA copy number.
- ◆ Main functionalities
  - ◆ Data pre-processing such as imputation of missing values using lowess approach and filtering.
  - ◆ Set of visualization functions for displaying measured and derived information as a function of genomic position.
  - ◆ Utilities to perform and interpret tests for associations between clinical variables and copy number of individual loci as well as collective features of genomic profiles.
  - ◆ Implementation of the HMM-based algorithm for finding genomic events e.g., copy number transitions and high-level amplifications.

## Data Processing / QA plots

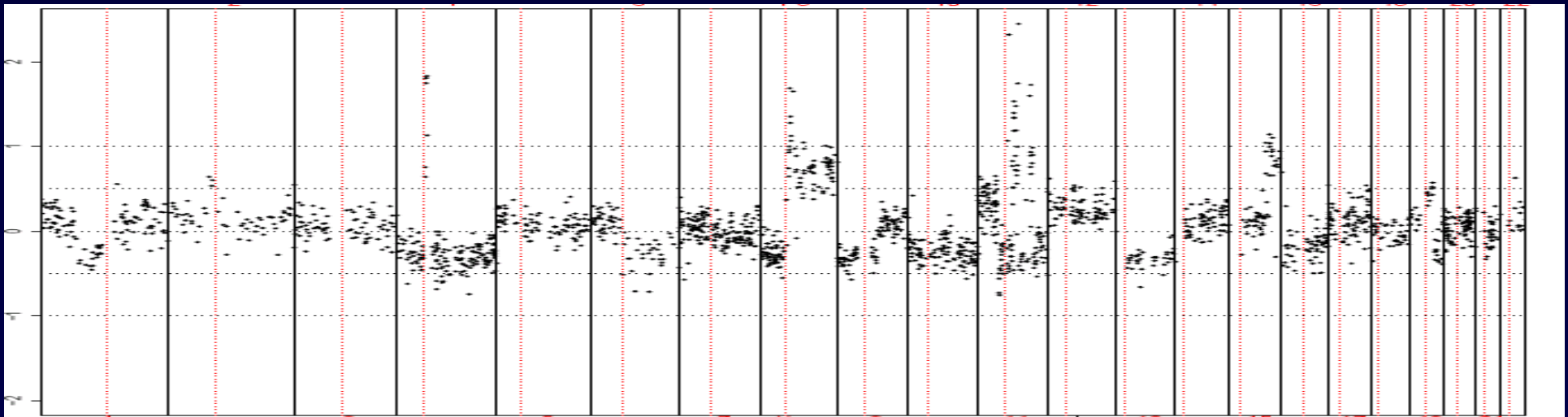


## Visualization Tools

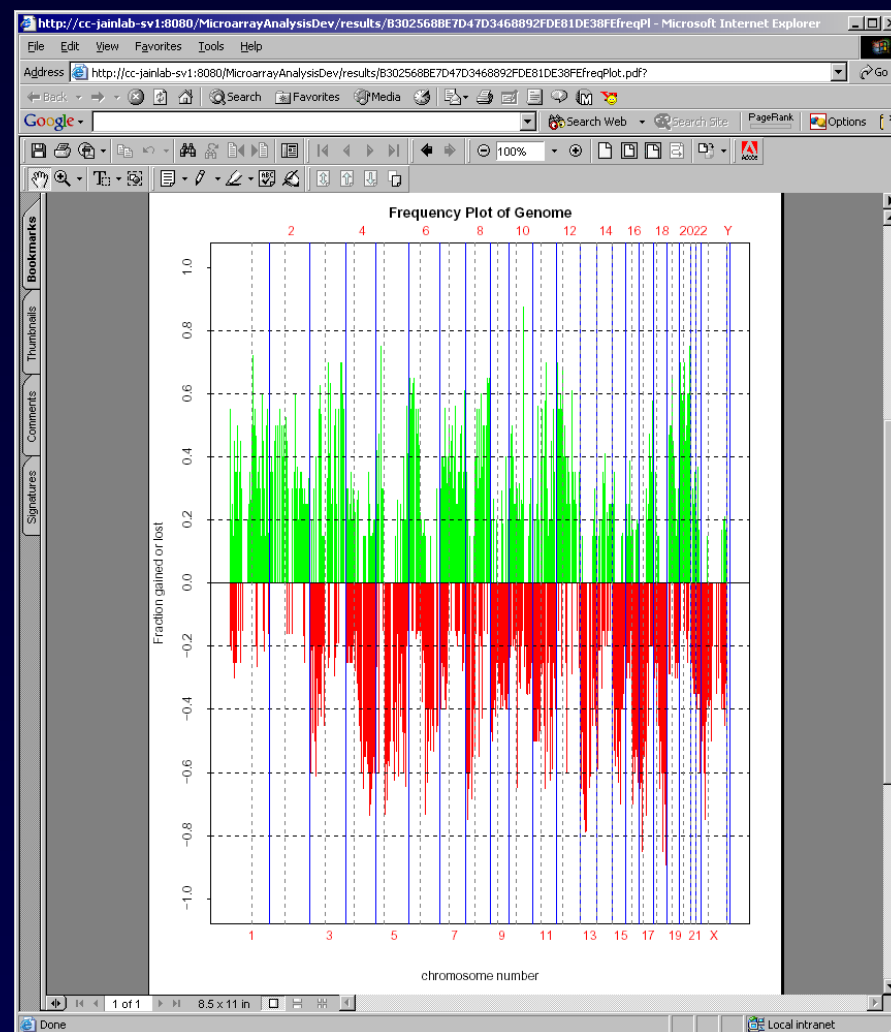
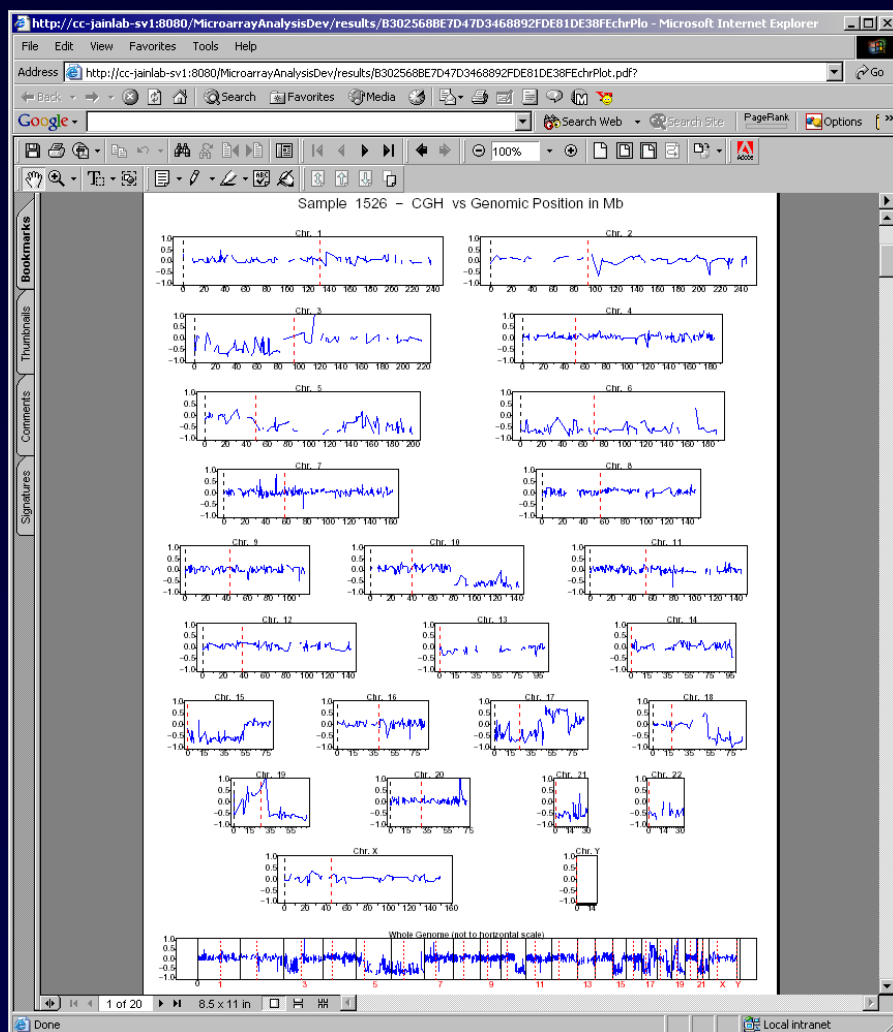
Frequency plot



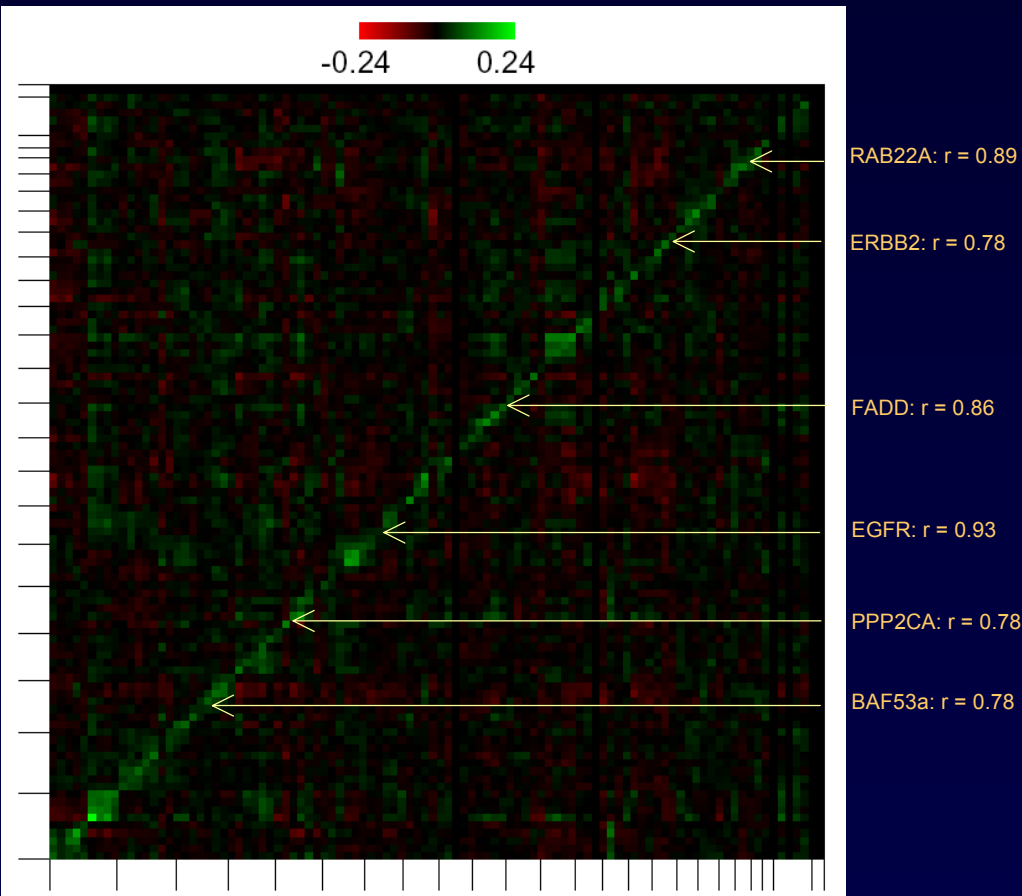
Genomic profile



## Visualization Tools – some already in Magellan



## Visualization Tools – effect of copy number on transcription

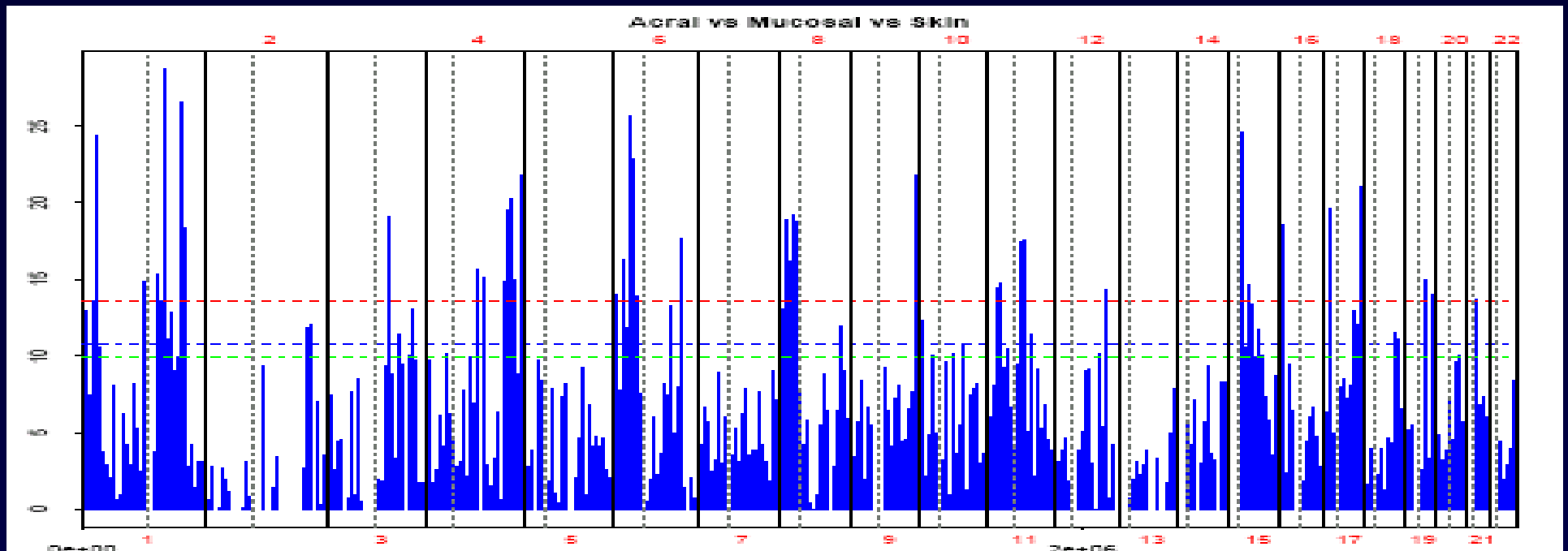


Microsoft Excel - BAC\_GeneCorrelations(1MB)\_annotated.txt

	A	B	C	D
	CorrPairs	BAC to Gene Distance(bp)	Corr Coefficient	Gene Name
1	CGH[176](199) vs. Expression[18123](218219_s_at)	237029	0.9738	LANCL2
2	CGH[177](200) vs. Expression[18123](218219_s_at)	237029	0.968325	LANCL2
3	CGH[176](199) vs. Expression[8139](208091_s_at)	282505	0.94593	DKFZP564K0822
4	CGH[177](200) vs. Expression[8139](208091_s_at)	282505	0.940003	DKFZP564K0822
5	CGH[176](199) vs. Expression[3550](203484_at)	432427	0.934059	SEC61G
6	CGH[176](199) vs. Expression[2051](201983_s_at)	17083	0.931767	EGFR
7	CGH[176](199) vs. Expression[10970](210984_x_at)	34146	0.928246	EGFR
8	CGH[177](200) vs. Expression[10970](210984_x_at)	34146	0.917171	EGFR
9	CGH[177](200) vs. Expression[3550](203484_at)	432427	0.9171	SEC61G
10	CGH[177](200) vs. Expression[2051](201983_s_at)	17083	0.914904	EGFR
11	CGH[176](199) vs. Expression[11551](211607_x_at)	34146	0.914642	
12	CGH[309](341) vs. Expression[11975](212050_at)	485815	0.913807	
13	CGH[239](265) vs. Expression[18005](218101_s_at)	857137	0.909432	NDUFC2
14	CGH[177](200) vs. Expression[11551](211607_x_at)	34146	0.90828	
15	CGH[313](347) vs. Expression[11974](212049_at)	107818	0.908238	
16	CGH[176](199) vs. Expression[2052](201984_s_at)	18508	0.906663	EGFR
17	CGH[309](341) vs. Expression[12632](212708_at)	430365	0.906324	
18	CGH[309](341) vs. Expression[11974](212049_at)	488059	0.904665	
19	CGH[309](341) vs. Expression[11976](212051_at)	488671	0.902152	
20	CGH[313](347) vs. Expression[11975](212050_at)	110062	0.899056	
21	CGH[282](312) vs. Expression[13296](213376_at)	553065	0.899045	
22	CGH[383](415) vs. Expression[18264](218360_at)	794307	0.894177	RAB22A
23	CGH[313](347) vs. Expression[11914](211988_at)	273955	0.890524	SMARCE1
24	CGH[177](200) vs. Expression[2052](201984_s_at)	18508	0.888031	EGFR
25	CGH[309](341) vs. Expression[11914](211988_at)	869833	0.887815	SMARCE1
26	CGH[309](341) vs. Expression[19136](219233_s_at)	150141	0.882241	PRO2521
27	CGH[239](265) vs. Expression[3611](203545_at)	828823	0.88017	MGC2840
28	CGH[306](335) vs. Expression[18662](218759_at)	361778	0.876803	DVL2
29	CGH[176](199) vs. Expression[5260](205194_at)	514452	0.873193	PSPH
30	CGH[309](341) vs. Expression[7900](207842_s_at)	406520	0.87312	MLN51
31	CGH[313](347) vs. Expression[11976](212051_at)	107206	0.872482	
32	CGH[176](199) vs. Expression[18123](218219_s_at)	237029	0.871973	LANCL2

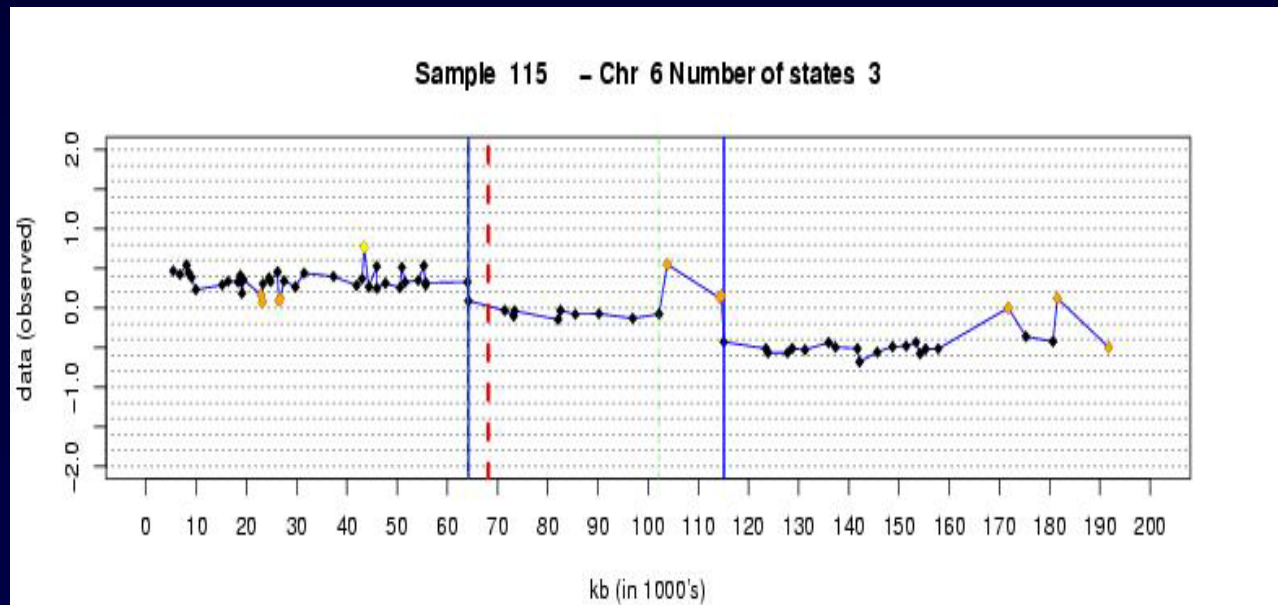
## Visualization of Statistics

Statistics and significance cut-off



## HMM based approach to automatically identify structural abnormalities in tumor genomes using array CGH data

- ◆ Fridlyand et al. Journal of Multivariate Analysis 90 (2004) 132-153



Questions?

caArray is currently intended as a repository for raw, not normalized data

- ◆ Is there a push toward implementing normalization and storage of processed data as part of caArray?
- ◆ Since several projects have integration of caArray / caBIO as a deliverable, are all of us going to have to individually provide normalization tools and store processed data locally in the short term?

April 1, 2005 2:00pm EST

## Discussion

- ◆ Project updates
- ◆ Any requests?